



# 군집분석(Clustering Analysis)

---

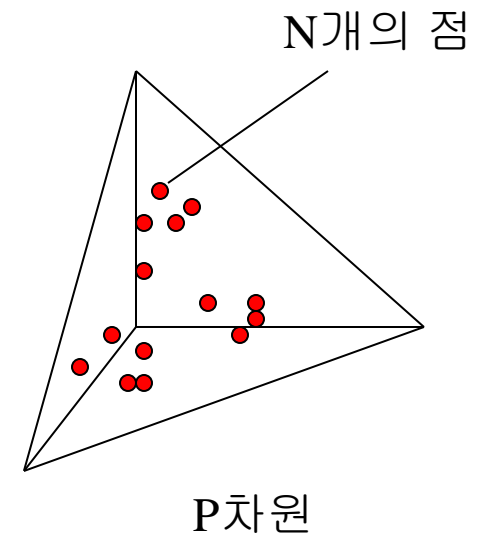
# 군집분석(Clustering Analysis) 개요

- 다변량 자료를 각 특성의 유사성에 따라 여러 그룹(군집 또는 집락)으로 나누는 통계적 기법중의 하나

P개

ID	성별	나이	직업	학력	본인소득	가입일	연횡수	연간금액	결제
1	1	21	7	2		9908	1	10000	1
3	2	25	2	4	3	9902	2	124000	1
4	2	25	2	4	3	9902	2	124000	1
5	1	30	2	4	7	9705	2	157000	1
6	2	47	4	1	5	9904	1	10000	1
8	1	24	7	2	2	9804	1	10000	1
14	2	22	2	4		9908	2	98000	1
15	2	22	2	4		9908	2	98000	1
16	1	19	7	2		9908	1	34500	1
17	1	37	5	4	5	9710	2	57000	1
18	1	37	5	4	5	9710	2	57000	1
19	1	39	2	6	2	9812	2	40500	1
20	1	39	2	6	2	9812	2	40500	1

N개



- 군집의 개수, 내용, 구조가 파악되지 않은 상태에서 특성을 파악하며, 군집들 간의 관계를 분석 (탐색적 분석)
- 고객의 세분화 또는 군집 별로 추가적인 분석을 수행하기 위해 활용

# 유사성 측정

## Clustering에서의 거리 계산

### 유사성

- 군집으로 묶기 위해서는 개체간에 유사한 특성을 가지고 있어야 함
- 이 유사한 특성의 정도를 나타내는 척도로 개체간의 거리를 사용하고, 거리가 상대적으로 가까운 개체들을 동일 군집으로 묶음

- 거리(distance)의 조건

$$d_{ij} \geq 0, d_{ii} = 0, i, j = 1, 2, \dots, n$$

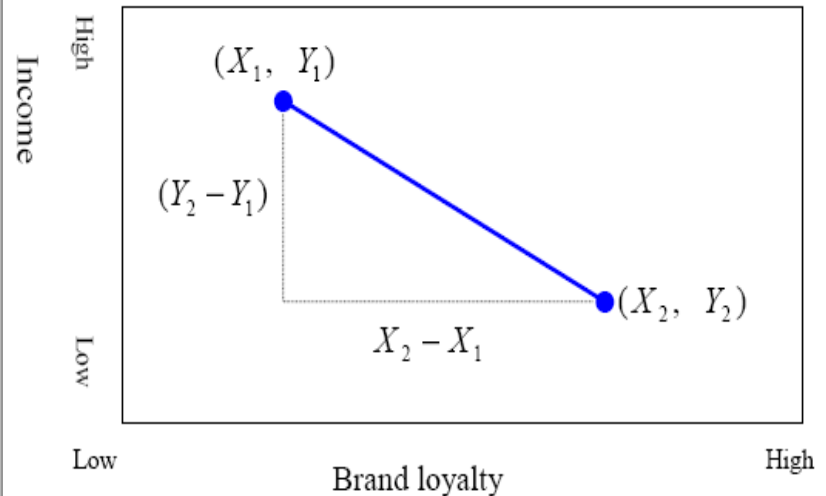
$$d_{ij} = d_{ji}, i, j = 1, 2, \dots, n$$

$$d_{ij} + d_{jk} \geq d_{ik}, i, j, k = 1, 2, \dots, n$$

- 개체간의 거리는 행렬을 이용하여 계산  
유클리드안 거리/유클리드안 제곱 거리/  
시티-블록, 맨하탄 거리/코사인 거리/체비셰프 거리/민코우스키 거리 등이 있음

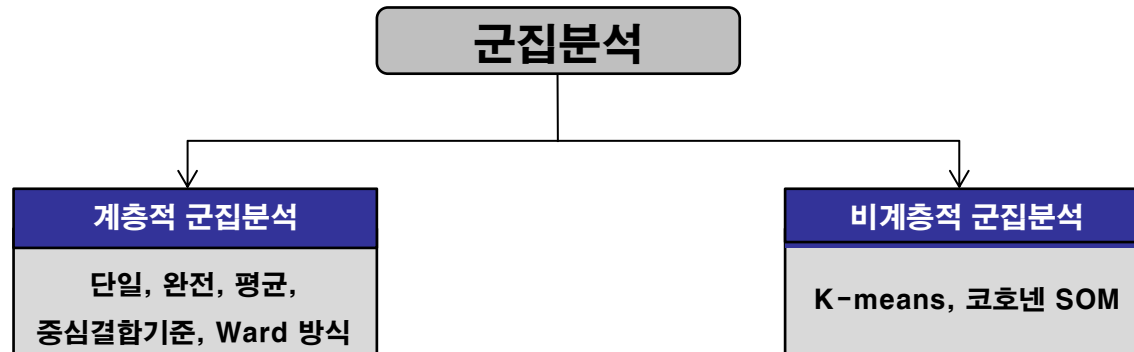
e.g.

### 유클리드 거리



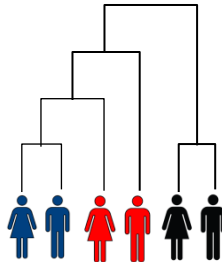
$$\text{Distance} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

# 군집분석의 종류



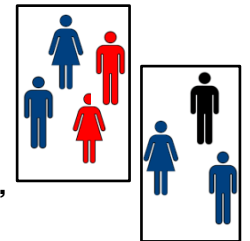
개별대상 간의 거리에 의하여 가장 가까이 있는 대상들로부터 시작하여 결합해 감으로써 나무모양의 계층구조를 형성해가는 방법

- 장점: 군집이 형성되는 과정을 정확하게 파악할 수 있어 군집의 수 도출이 용이
- 단점: 자료의 크기가 크면 분석하기 어려움.



군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함해 가는 방식으로 군집을 형성하는 방법

- 장점: 많은 자료를 빠르고 쉽게 분류
- 단점: 군집의 수를 미리 정해 주어야 하고, 군집을 형성하기 위한 초기값에 따라 군집결과가 달라짐.



# 군집분석의 활용 (1/5)

산업(대)	산업(중)	주제(상황)	변수	결과
유통	백화점	<ul style="list-style-type: none"> <li>쇼핑 성향을 통해 고객들을 군집화하고자 함.</li> </ul>	<ul style="list-style-type: none"> <li>[7점 척도] 쇼핑은 흥미 있음, 쇼핑은 당신의 소득에 악영향을 끼침, 쇼핑을 하면서 외식을 즐기, 쇼핑 시 최고 제품을 구입하기 위한 노력, 쇼핑에 관심이 없음.</li> </ul>	<p>3개의 군집으로 도출됨.</p> <p>군집1. 쇼핑의 흥미, 쇼핑을 하면서 외식을 즐길의 평균이 높음. 쇼핑에 관심이 없음은 평균이 낮음. → 쇼핑 애호가 군</p> <p>군집2. 쇼핑의 흥미, 쇼핑을 하면서 외식을 즐길의 평균이 낮음. 쇼핑에 관심이 없음의 평균이 높음. → 냉담한 소비자 군</p> <p>군집3. 쇼핑은 가계에 악영향, 쇼핑 시 최고의 상품을 구입하기 위한 노력의 평균이 높음. → 경제적인 소비자 군</p>
서비스	호텔	<ul style="list-style-type: none"> <li>호텔 종사원의 특성을 분류하고자 함.</li> </ul>	<ul style="list-style-type: none"> <li>사회적 책임활동, 조직몰입, 근속연수, 연령, 학력, 성별, 결혼여부</li> </ul>	<p>2개의 군집으로 도출됨.</p> <p>군집1. 조직몰입, 근속연수, 연령의 평균이 높음. 군집2. 조직몰입은 높으나 근속연수, 연령의 평균이 낮음.</p>

# 군집분석의 활용 (2/5)

산업(대)	산업(중)	주제(상황)	변수	결과
유통	백화점	<ul style="list-style-type: none"> <li>고객 등급화</li> </ul>	<ul style="list-style-type: none"> <li>나이, 성별, 주소, 주거형태, 집 평수, 백화점 첫 이용 날짜, 구매일자, 항목, 구매액수, 결제수단, 첫 구매 시기</li> </ul>	<ol style="list-style-type: none"> <li>A백화점의 고객은 4개의 등급으로 분류됨.</li> <li>기존고객들의 등급 별 특성을 도출 후 신규고객을 유입하기 위한 방안을 모색함.</li> </ol>
서비스	골프장	<ul style="list-style-type: none"> <li>만족 유형을 이용한 집단 분류</li> </ul>	<ul style="list-style-type: none"> <li>[골프 연습장의 만족척도] 시설, 요금, 대인서비스의 요인을 요인점수로 환산</li> </ul>	<p>5개의 군집으로 도출됨.</p> <ol style="list-style-type: none"> <li>시설 만족군</li> <li>전반적 만족군</li> <li>비용 만족군</li> <li>대인서비스 만족군</li> <li>전반적 불만족군</li> </ol>
	커피 전문점	<ul style="list-style-type: none"> <li>이용실태를 분석하여 집단 분류</li> </ul>	<ul style="list-style-type: none"> <li>테이크아웃 전문점 이용횟수, 이용 목적, 구입한 음식의 용도, 1회 평균 지출액</li> </ul>	<p>2개의 군집 중 군집1은 메뉴의 다양성, 매장 기기 및 기물의 청결성 속성에 높은 중요도를 나타냄. -&gt; 다양한 메뉴 개발, 청결 서비스 전략을 수행해야 함.</p>

# 군집분석의 활용 (3/5)

산업(대)	산업(중)	주제(상황)	변수	결과
공공	군수산업	<ul style="list-style-type: none"> <li>여군의 새로운 군복 치수 결정</li> </ul>	<ul style="list-style-type: none"> <li>가슴, 목, 어깨둘레, 소매 바깥솔기, 목에서 엉덩이까지의 길이 등</li> </ul>	20가지의 형태로 구성된 의복치수 군집이 도출됨
제조	목욕세제	<ul style="list-style-type: none"> <li>마케팅전략을 도출하기 위해 고객을 분류하고자 함.</li> </ul>	<ul style="list-style-type: none"> <li>묶음판매 프로모션 구매 여부, 구매한 브랜드 수, 연속해서 구매한 브랜드 수, 구매거래 수, 인구통계량 정보</li> </ul>	<p>4개의 군집으로 도출됨</p> <p><b>군집1.</b> 구매가 가장 많이 일어나는 집단이긴 하나 구매하는 브랜드 수나 연속해서 구매하는 브랜드 건수가 많음. → 브랜드에 대한 충성도가 높은 집단으로 보긴 어려움.</p> <p><b>군집2.</b> 프로모션 반응을 가장 낮음.</p> <p><b>군집3.</b> 구매거래의 수에 비해 브랜드의 수가 적고 연속해서 구매한 브랜드의 건수가 많이 나타남.</p> <p><b>군집4.</b> 35세 이상의 여성이 주를 이룸.</p>

# 군집분석의 활용 (4/5)

산업(대)	산업(중)	주제(상황)	변수	결과
서비스	온라인 게임	<ul style="list-style-type: none"> <li>고객의 행위 특성을 분류</li> </ul>	<ul style="list-style-type: none"> <li>총 사용 시간, 총 사용 횟수, 에러 횟수, 한달 평균 사용시간</li> </ul>	<p>그리드 5X7의 군집이 가장 적합함.</p> <p>각 군집의 특성별로 행위 유도 방안을 제시함. (접속횟수의 증가, 평균 사용시간의 증가, 총 사용시간의 증가, 접속횟수 및 총 사용시간의 증가, 사용 오류의 감소)</p>
금융	은행	<ul style="list-style-type: none"> <li>주택담보대출을 받은 고객의 특성을 파악하기 위해 분류</li> </ul>	<ul style="list-style-type: none"> <li>집의 감정가, 사용 가능한 신용 잔액, 주어진 대출금액, 연령, 결혼 상태, 자녀의 수, 가구 소득, 입출금 시스템, 신용카드 시스템</li> </ul>	<p>주택담보대출을 받은 고객의 군집엔 [대학에 진학하는 자녀들을 둔 고객층]에 많은 고객이 포함, 그들은 개인 계좌뿐 아니라 기업 계좌도 보유하고 있음.</p> <p>→ 자녀들이 대학을 가서 집을 떠나면 대출금을 이용하여 새로운 사업을 시작할 기회를 엿봄.</p> <p>→ 이를 통해 은행은 집의 빈자리를 이용한 사업을 꾸려나갈 부모들을 목표로 하는 새로운 마케팅 프로그램을 생성함.</p>



# 군집분석의 활용 (5/5)

산업(대)	산업(중)	주제(상황)	변수	결과
금융	은행	<ul style="list-style-type: none"> <li>인터넷 뱅킹 고객의 특성요인을 통해 고객을 군집화하고자 함.</li> </ul>	<ul style="list-style-type: none"> <li>유용성, 사용편의성, 신뢰성, 위험성, 실제사용</li> </ul>	<p>4개의 군집으로 도출됨.</p> <p>군집1. 편의성 요인을 높일 수 있는 마케팅 전략이 필요함.</p> <p>군집2. 타 군집군에 비해 인터넷 뱅킹 이용률이 매우 저조함 → 충성도를 높이는 전략이 필요함.</p> <p>군집3. 자사의 인터넷 뱅킹 이용이 안전하다는 인식을 지속적으로 알려 주는 마케팅 전략이 필요함.</p> <p>군집4. 핵심 우량 고객군 → 차별화된 고객관리가 필요함. (ex. 은행 수수료 감면 or 다양한 개인별 금융정보 서비스를 실시할 수 있음)</p>
	증권	<ul style="list-style-type: none"> <li>균형 포트폴리오를 구성하기 위해 투자대상 기업을 분류</li> </ul>	<ul style="list-style-type: none"> <li>수익(일별, 주별 또는 월별), 가격변동률, 베타, 자본총액 등</li> </ul>	<p>도출된 서로 다른 군집으로부터 주식을 선택하여 위험을 분산</p>



# 군집분석의 평가

---

- 군집분석에는 분석 전에 정해야 하는 사항이 많다 (예: 초기 군집수, 가중치, 거리측도 등)
- 분석자의 주관에 의하여 결정되는 이러한 사항들이 군집분석의 결과에 어떻게 영향을 미치는 가를 알아보기 위하여는, 군집분석 결과의 평가가 필수적이다.
- 좋은 결과는 각 군집 안에서의 분산이 최소로 되는 것이다.
- 또는, 사용되어진 거리의 측도를 이용하여 군집내의 거리의 평균과 군집간의 거리의 평균을 비교할 수 있다. 즉, 군집내의 거리의 평균이 군집간의 거리의 평균 보다 작으면 좋은 결과이다.



# 군집분석의 장점

---

- 비통제 지식 발견 기술이다.
  - 군집분석은 그 자체가 대용량 데이터에 대한 탐색적인 기법으로서, 주어진 데이터의 내부구조에 대한 사전적인 정보없이 의미있는 자료구조를 찾아낼 수 있다.
- 다양한 형태의 데이터에 적용가능
  - 거리만 잘 정의되면, 모든 종류의 자료에 적용할 수 있다. 예를 들면, 신문기사와 같은 텍스트 자료도 그 거리만 잘 정의하면 얼마든지 군집분석을 사용할 수 있다.
- 분석방법의 적용 용이성
  - 자료의 사전정보를 필요로 하지 않아서 누구나 쉽게 분석할 수 있다.



# 군집분석의 단점

---

- 가중치와 거리 정의
  - 가중치와 거리를 어떻게 정의하는가에 따라 군집분석의 결과가 아주 민감하게 반응한다.
- 군집 수 결정이 쉽지 않다.
  - 사전에 정의된 군집 수를 기준으로 동일한 수의 군집을 찾게 되므로 만일 군집수가 원 데이터 구조에 적합하지 않다면 좋은 결과를 얻을 수 없다.
- 결과의 해석이 어렵다.
  - 찾아진 군집이 무엇을 의미 하는지 데이터만을 이용해서는 알 수가 없다.

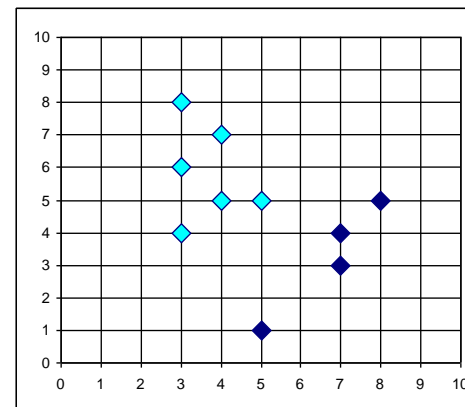
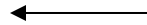
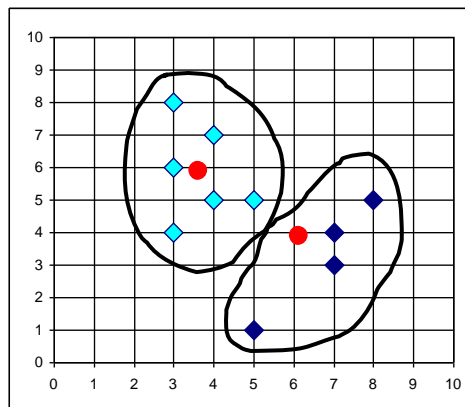
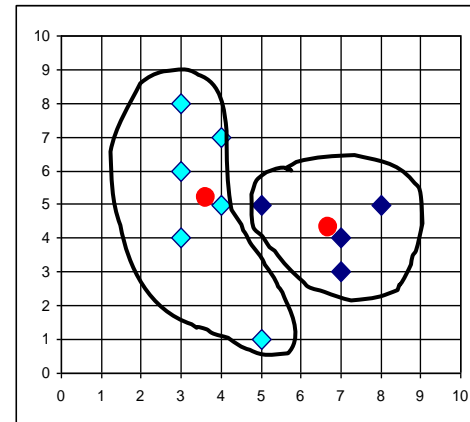
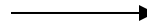
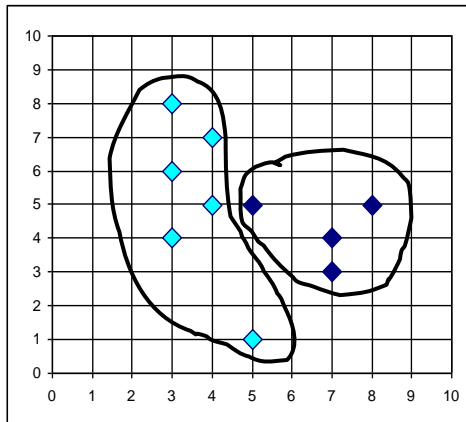


# K-평균 군집분석(K-Means algorithm)

---

- 단계 0 : 사전적으로 군집의 수  $K$ 를 지정한다.
- 단계 1 : 각 군집에 1개의 군집 중심을 임의로 정한다.  
(보통 서로 상당히 떨어진 개체를 선택함)
- 단계 2 : 모든 개체를 각각 가장 가까운 군집 중심에 배속시킨다.
- 단계 3 : 각 군집의 중심을 산출한다.
- 단계 4 : 단계 2와 단계3을 변화가 거의 없을 때까지  
(보통 10회 이하) 반복한다.

# K-평균 군집분석



# K-평균 군집분석 예제

## ▶ 1단계 최초 군집화

개체	변수1	변수2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5
8	6.0	2.0
9	5.0	3.0
10	6.5	3.0
11	7.0	3.5

최초 임의의 중심점 생성

개체	군집1의 중심(개체1)과의 거리( $\sqrt{}$ )	군집2의 중심(개체4)과의 거리( $\sqrt{}$ )	군집3의 중심(개체8)과의 거리( $\sqrt{}$ )
2	1.25	37.25	20.25
3	13	13	13
5	22.25	6.25	15.25
6	28.25	4.25	11.25
7	18.25	9.25	12.5
9	28.25	6.25	3.25
10	34.25	18.25	1.25
11	42.25	16.25	3.25

$$(1-1.5)^2 + (1-2)^2 = 1.25$$

최초 임의의 중심점과 개체간의 거리 계산

군집	개체	변수1	변수2	평균 (변수1, 변수2)	
1	1	1.0	1.0	1.84	2.34
	2	1.5	2.0		
	3	3.0	4.0		
2	4	5.0	7.0	4.125	5.375
	5	3.5	5.0		
	6	4.5	5.0		
	7	3.5	4.5		
3	8	6.0	2.0	6.125	2.88
	9	5.0	3.0		
	10	6.5	3.0		
	11	7.0	3.5		

최초 발생한 그룹들의 평균을 생성

## ▶ 2단계 반복

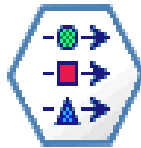
군집중심이 (1,1) → (1.84, 2.34)로 이동하였고, 이 중심을 기준으로 다시 군집간 개체 거리 계산을 하여, 변화가 없거나, 특정한 반복 수만큼 수행하고 멈추며, 멈추는 그 시점의 평균값이 곧 군집 중심이 되며, 그 중심으로부터의 거리가 최종 유사성 척도가 됨.

# Clementine에서의 K-평균 군집화

- K-평균 군집화에 필요한 노드 3가지



Var. File



Type



K-Means

-  의 자료 해석

K-Means

-  를 이용해 필요한 자료 도출

K-Means



# K-Means 노드



K-Means

K-Means

Fields Model Expert Annotations

Model name: ☐ Auto ☐ Custom

☒ Use partitioned data

Number of clusters:  군집 수 결정

☐ Generate distance field

Cluster label: ☒ String ☐ Number

Label prefix:

Optimize: ☐ Speed ☒ Memory

군집 중심으로부터의 거리  
(이상치 발견에도 종종 사용함)

OK Run Cancel Apply Reset

K-Means

Fields Model Expert Annotations

Mode: ☐ Simple ☒ Expert

Stop on: ☐ Default ☒ Custom

Maximum iterations:

Change tolerance:

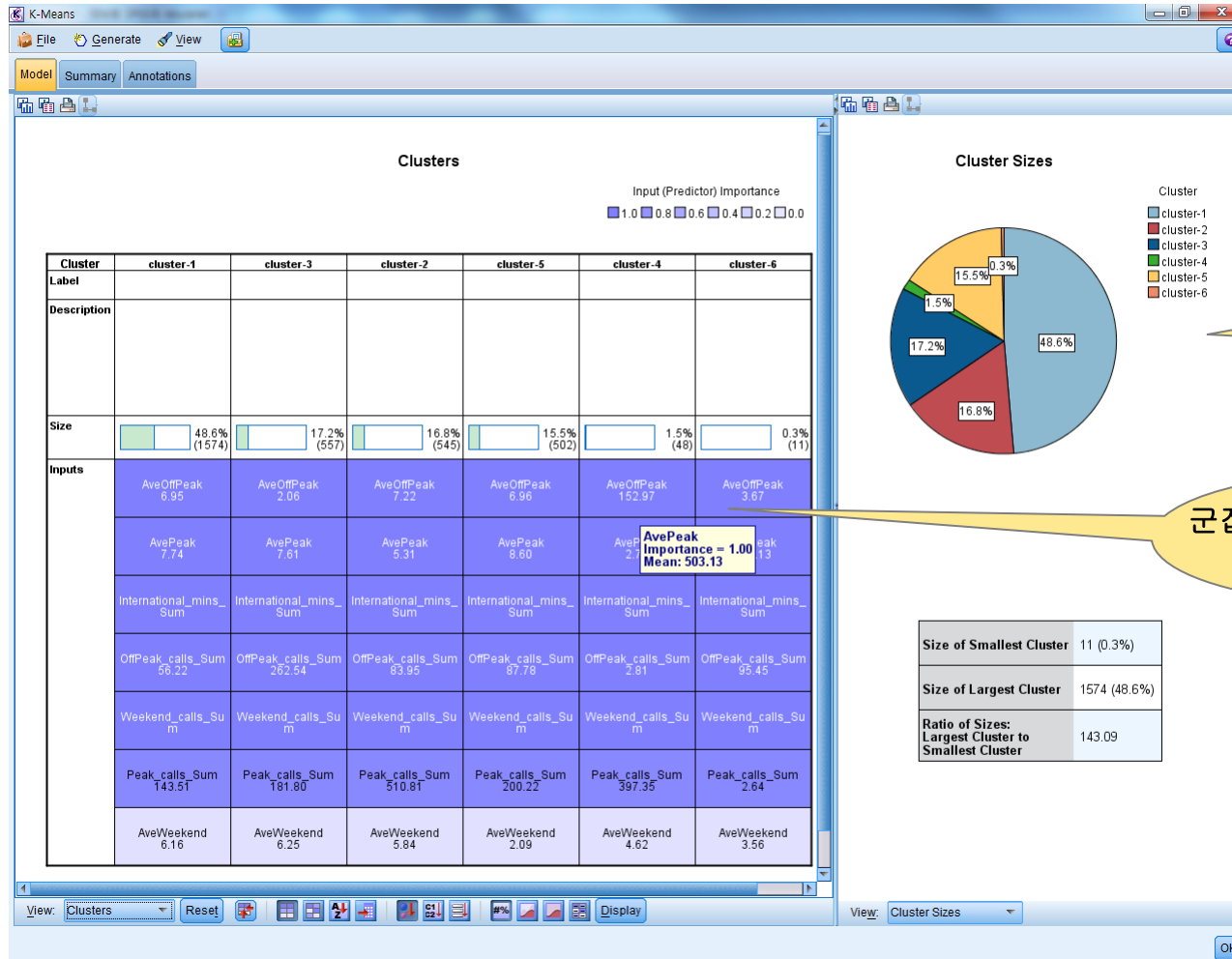
Encoding value sets:

최대 반복수 지정

군집 중심 변화량 기준 설정

OK Run Cancel Apply Reset

# K-Means 노드: 모델링 결과

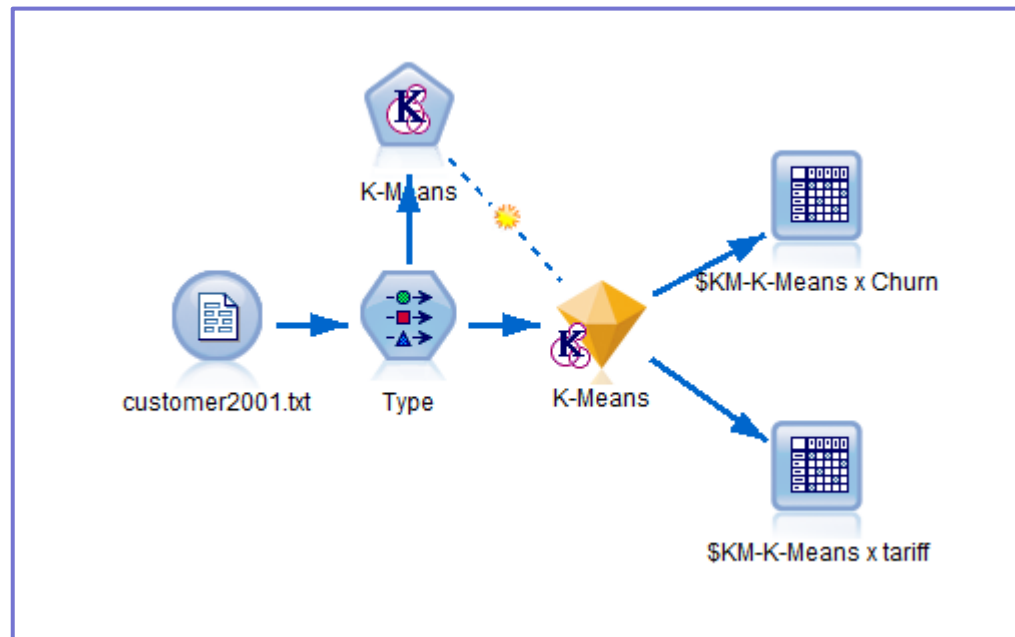


각 군집의 빈도를  
원도표로 표현

군집별로 변수의 중심값을  
비교할 수 있음

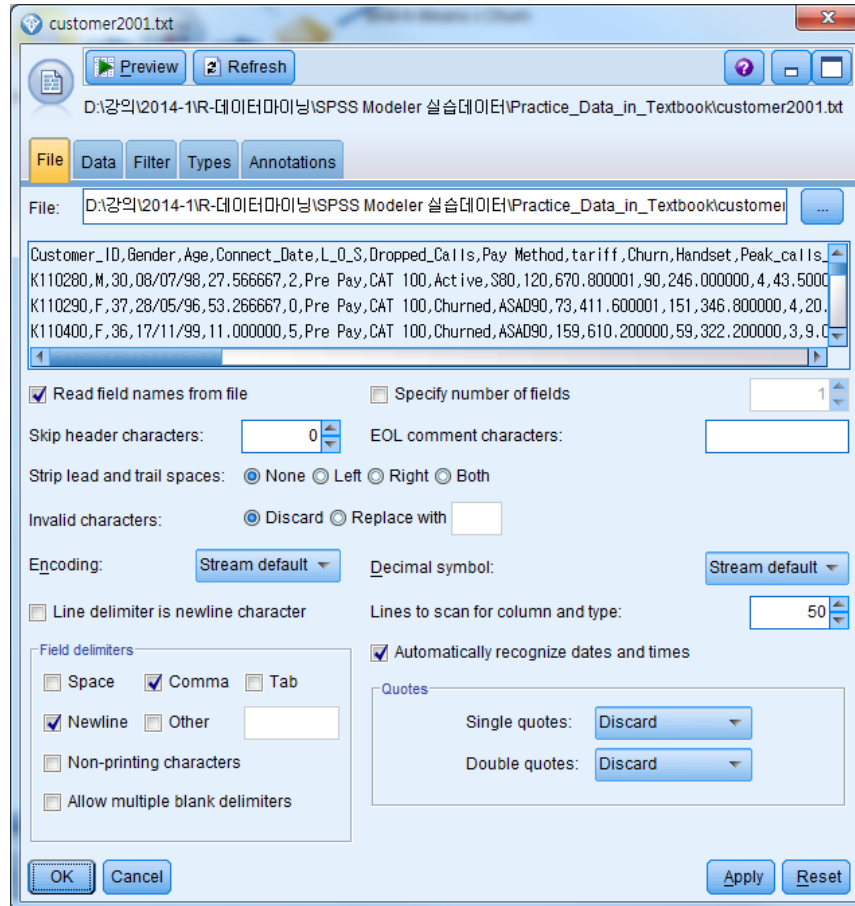
# K-평균 군집분석 실습 스트림

통신회사 고객 데이터(customer2001.txt)로 ID번호, 성, 나이 등의 신원 및 인구 정보, 이탈여부(Churn), 요금제(Tariff) 등의 계좌 기록, 주간 통화(Peak), 야간 통화(Off-Peak), 주말 통화(Weekend) 관련 통화 횟수 및 통화시간 등의 통화 정보를 포함하여 총 34개의 변수로 구성되어 있다. 이 중 일부 변수만을 선택하여 군집분석을 실시해 보기로 한다.



# Step1. 데이터 설명

## Clementine>K-means Clustering

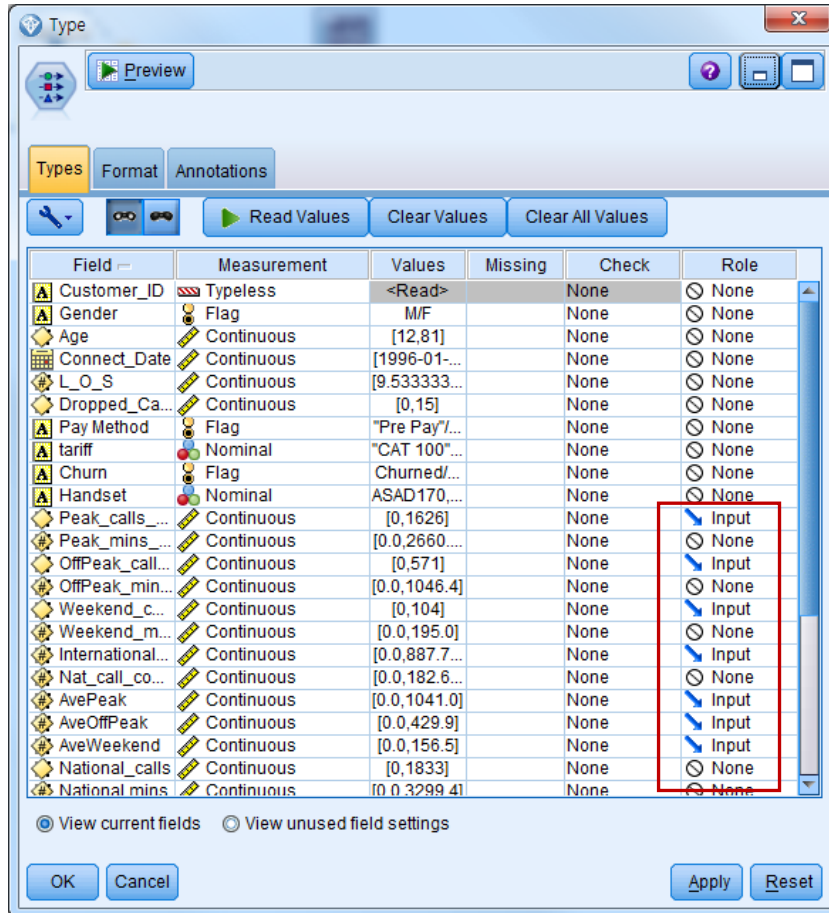


### ▶ 데이터 읽어오기

- Source- Variable File  
읽어올 데이터 경로 설정
- Read field names from file 클릭  
변수명이 데이터의 첫줄에 있음
- Delimiters  
구분자 설정(Comma, Newline)

# Step2. 변수 설정

## Clementine>K-means Clustering

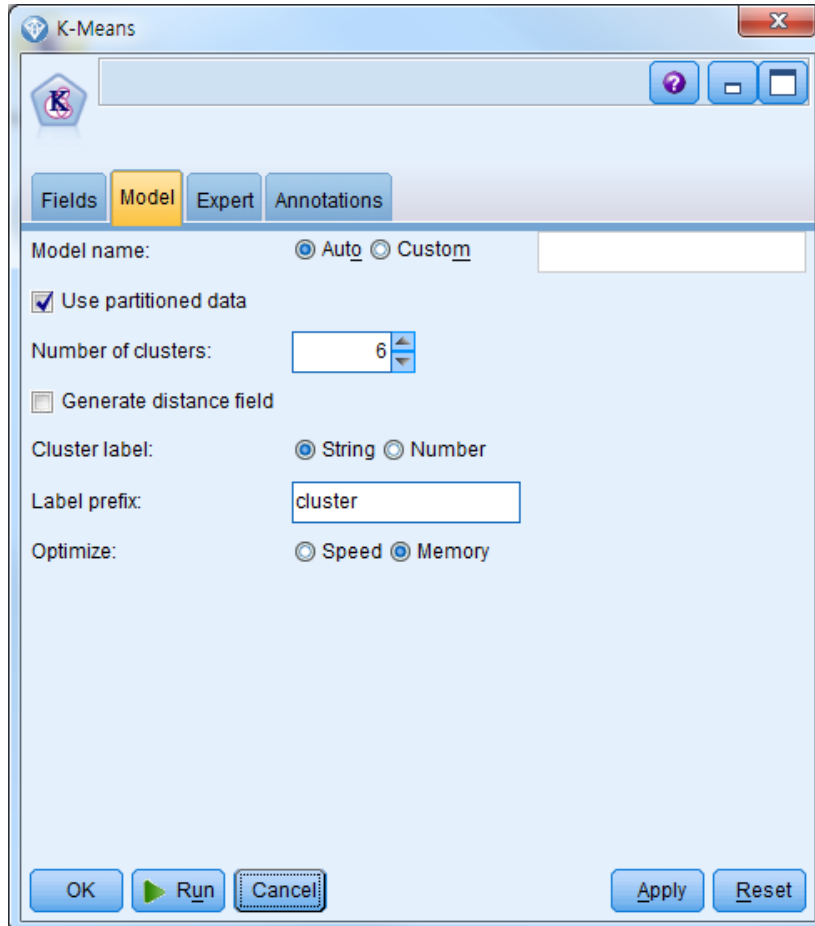


### ▶ 군집화 변수 지정

- 고객의 통화 관련 행태(behavior)와 관련된 7개 변수 군집화
  - Peak\_calls\_Sum(주간 총 통화횟수)
  - AvePeak(주간 평균 통화시간(분), 통화 1건당)
  - OffPeak\_calls\_Sum(야간 총 통화횟수)
  - AveOffPeak(야간 평균 통화시간(분), 통화 1건당)
  - Weekend\_calls\_Sum(주말 총 통화횟수)
  - AveWeekend(주말 평균 통화시간(분), 통화 1건당)
  - International\_min\_Sum(국제통화 총 통화시간(분))
- Field - Type
- Direction에서 군집화 변수 7개만 In으로 놓고 나머지 변수들은 None으로 둔다.
- 군집화 변수들의 유형은 모두 Continuous로 둔다.

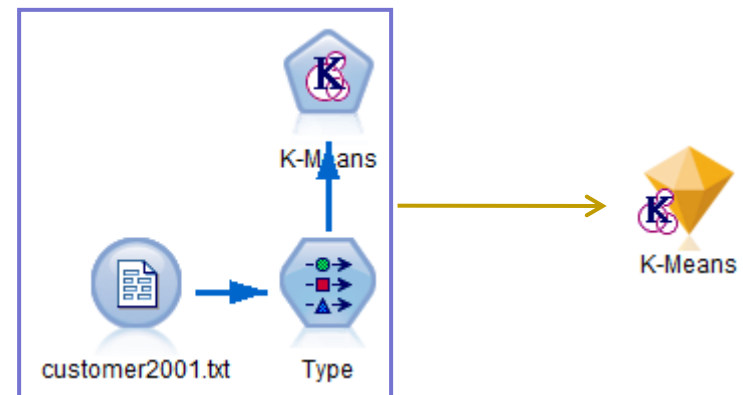
# Step3. 모델링 및 결과 (1/3)

## Clementine>K-means Clustering



### ▶ K-means clustering

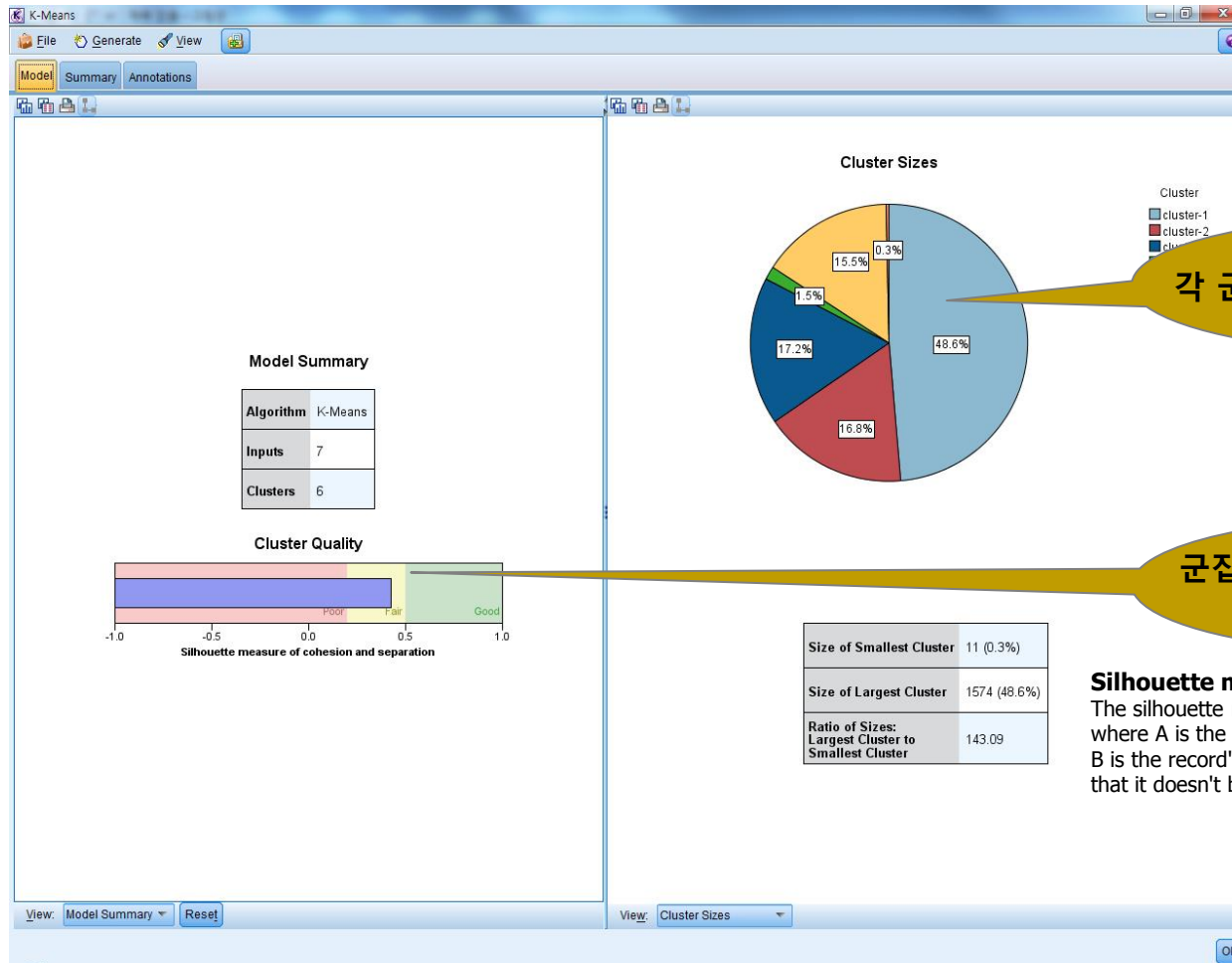
- Modeling –K-means
- 군집수가 6개인 K-평균 군집분석 실시
- Number of clusters 칸에 6 입력
- 실행(Execute)하면 다음과 같은 모델링이 생성됨



# Step3. 모델링 및 결과 (2/3)

## Clementine>K-means Clustering

▶ 모델링 결과



각 군집의 빈도를 원도표로 표현

군집분석의 성능을 시각적으로 확인할 수 있음

### Silhouette measure of cluster cohesion and separation

The silhouette measure averages, over all records,  $(B-A) / \max(A,B)$ , where A is the record's distance to its cluster center and B is the record's distance to the nearest cluster center that it doesn't belong to.

# Step3. 모델링 및 결과 (3/3)

## Clementine>K-means Clustering

Input (Predictor) Importance  
 1.0 0.8 0.6 0.4 0.2 0.0

### ▶ 모델링 결과

Cluster	cluster-1	cluster-2	cluster-3	cluster-4	cluster-5	cluster-6
Label						
Description						
Size	48.6% (1574)	16.8% (545)	17.2% (557)	1.5% (48)	15.5% (502)	0.3% (11)
Inputs	      	      	      	      	      	      

- **군집 1**  
주간 통화횟수, 주말 통화횟수, 국제통화시간 등이 큼  
-> 고빈도 가입자 그룹
- **군집 2**  
주간 통화횟수와 국제통화시간 등이 큰 가입자 그룹
- **군집 3**  
야간 통화횟수가 많고 주말 통화시간이 긴 가입자 그룹
- **군집 4**  
야간 통화횟수가 많은 가입자 그룹
- **군집 5**  
주말 통화횟수가 많은 가입자 그룹
- **군집 6**  
통화횟수가 매우 적지만 1통화당 통화시간이 매우 긴 가입자 그룹



# Step4. 모델 검증

## Clementine>K-means Clustering

\$KM-K-Means x Churn

Settings Appearance Output Annotations

Fields: ☒ Selected ☐ All flags (true values) ☐ All Numerics

Rows: \$KM-K-Means

Columns: Churn

☒ Include missing values

Cell contents: ☒ Cross-tabulations ☐ Function

Field:

Function: ☒ Mean ☐ Sum ☐ SDev ☐ Max ☐ Min

OK Run Cancel Apply Reset

\$KM-K-Means x tariff

Settings Appearance Output Annotations

Fields: ☒ Selected ☐ All flags (true values) ☐ All Numerics

Rows: \$KM-K-Means

Columns: tariff

☒ Include missing values

Cell contents: ☒ Cross-tabulations ☐ Function

Field:

Function: ☒ Mean ☐ Sum ☐ SDev ☐ Max ☐ Min

OK Run Cancel Apply Reset

Matrix of \$KM-K-Means by Churn

File Edit Genera

Matrix Appearance Annotations

Churn

\$KM-K-Means		Active	Churned
cluster-1	Count	809	765
	Row %	51.398	48.602
cluster-2	Count	274	271
	Row %	50.275	49.725
cluster-3	Count	332	225
	Row %	59.605	40.395
cluster-4	Count	21	27
	Row %	43.750	56.250
cluster-5	Count	282	220
	Row %	56.175	43.825
cluster-6	Count	7	4
	Row %	63.636	36.364

Cells contain: cross-tabulation of fields (including mis...)

Chi-square = 17.085, df = 5, probability = 0.004

OK

Matrix of \$KM-K-Means by tariff #1

File Edit Generate

Matrix Appearance Annotations

tariff

\$KM-K-Means		CAT 100	CAT 200	CAT 50	Play 100	Play 300
cluster-1	Count	488	373	204	410	99
	Row %	31.004	23.698	12.961	26.048	6.290
cluster-2	Count	24	507	0	0	14
	Row %	4.404	93.028	0.000	0.000	2.569
cluster-3	Count	44	231	0	84	198
	Row %	7.899	41.472	0.000	15.081	35.548
cluster-4	Count	6	31	1	6	4
	Row %	12.500	64.583	2.083	12.500	8.333
cluster-5	Count	124	187	19	101	71
	Row %	24.701	37.251	3.785	20.120	14.143
cluster-6	Count	1	10	0	0	0
	Row %	9.091	90.909	0.000	0.000	0.000

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 1,351.971, df = 20, probability = 0

OK

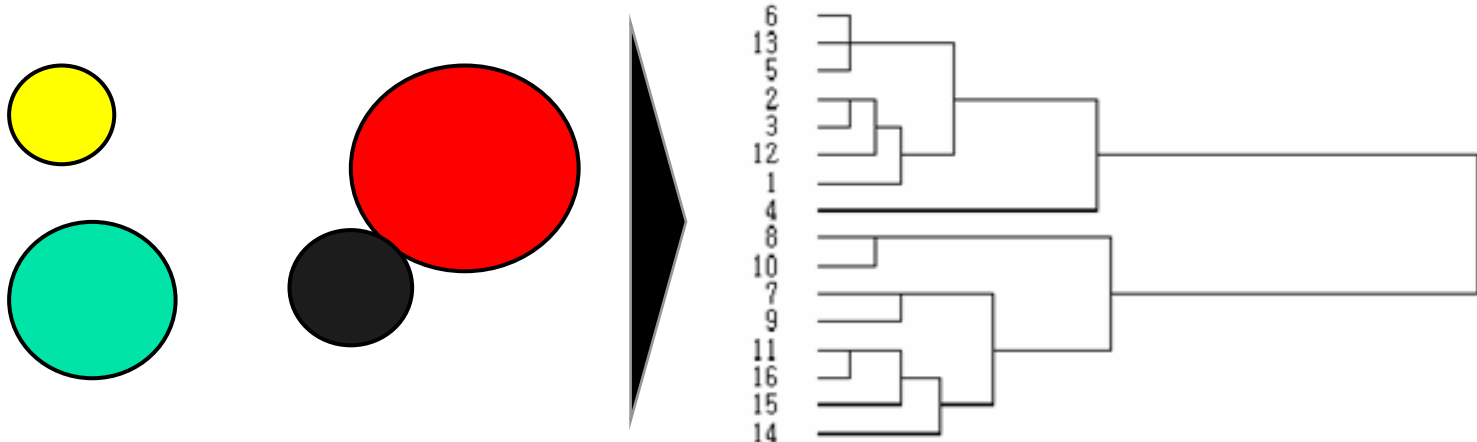
### ▶ 군집의 외적 특성

- 군집화에 사용하지 않았던 다른 변수와 교차표를 그려봄으로써 생성된 군집들의 여러 특성을 파악할 수 있음
- Output- Matrix
- 이탈(Churn) \* 군집분류(\$KM-Kmeans)  
군집 4의 이탈율이 다른 군집에 비해 비교적 높음
- 요금제(Tariff)\* 군집분류(\$KM-Kmeans)  
군집 2와 군집 6이 다른 군집에 비해 비교적 CAT 200 요금제를 많이 사용함

# 2단계 군집분석

계층적 군집분석과 비계층적 군집분석의 단점을 보완, 대용량 데이터의 군집화에 매우 효율적인 방법

TwoStep Method	
Step 1	<b>사전 군집화 (pre-cluster)</b> 순차적으로 개체들을 사전에 정한 수만큼의 예비-군집(sub-cluster)으로 만드는 과정: BIRCH 알고리즘 사용
Step 2	<b>계층적 군집화 (hierachical clustering)</b> 계층적 군집화 방법으로 전 단계의 예비 군집들을 원하는 군집 수만큼 주어진 조건을 만족할 때까지 군집화
거리측정	<ul style="list-style-type: none"> <li>2단계 군집분석에서는 개체간 (또는 군집간) 거리 측정에 '유클리디안(Euclidean)' 또는 '로그-우도(Log-likelihood)' 거리를 이용</li> <li>로그-우도 거리의 경우 연속형 변수는 정규분포를, 범주형 변수는 다항분포를 따르며 각각 변수들은 상호 독립인 것으로 가정</li> </ul>
군집 수 결정	2단계 군집분석에서는 계층적 군집화 방법을 이용하여 각각의 군집의 수에서 Schwarz's Bayesian Criterion(BIC) 또는 Akaike's Information Criterion(AIC) 통계량을 활용하여 최적의 군집 수를 결정

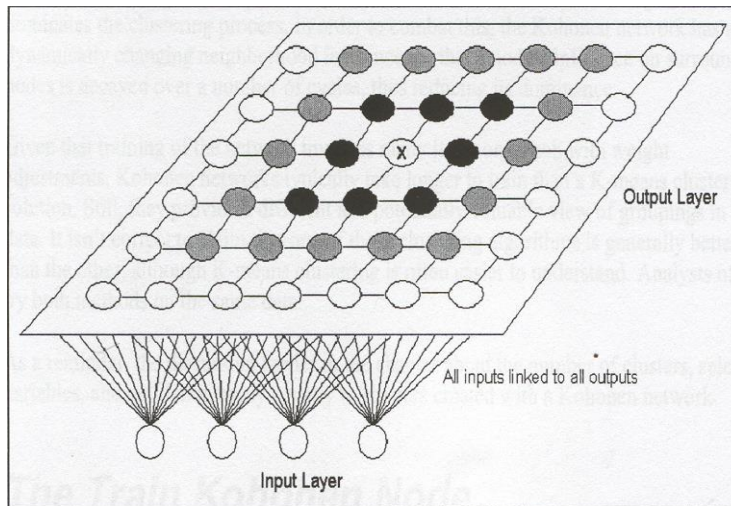


# Kohonen SOM 군집분석 (1/3)

SOM (Self-Organizing Map)은 자기조직화 지도라는 것으로 관측개체들을 스스로 조직화하여 지도의 형태로 뿌려주는 신경망 기법이다. SOM (자기조직화 지도)개념도는 2개의 층으로 이루어져 있으며, 첫 번째 층이 입력 층(input layer), 두 번째 층은 출력 층(output layer)으로 이루어진 2차원 격자(grid)로 되어 있다.

## Kohonen SOM Clustering

### Kohonen SOM의 개념도



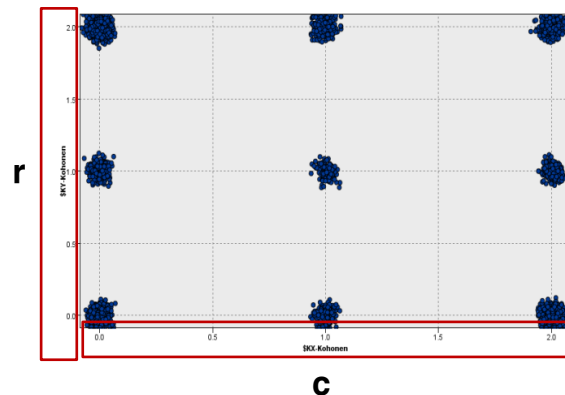
### Kohonen SOM의 사전작업

- SOM에 사용할 변수를 미리 지정해야 한다.
- 사용 변수들은 주 알고리즘의 적용에 앞서 표준화 되어야 한다.
- 2차원 SOM의 경우 그리드의 크기를 미리 결정한다.

# Kohonen SOM 군집분석 (2/3)

## Kohonen Method

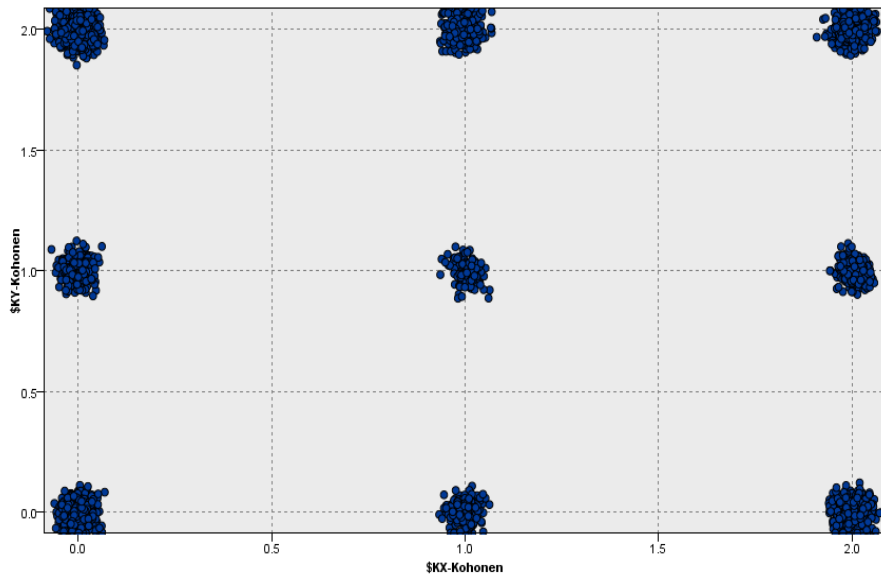
<b>Step 1</b>	2차원 그리드 상에 출력 노드( $r \times c$ )들을 깔아 놓는다. [initialization]디폴트 $10 \times 7$
<b>Step 2</b>	입력벡터(입력노드)를 표준화(Standardization) 한다. (범위: 0-1)
<b>Step 3</b>	각 레코드는 가장 유사한 출력노드, 중량 벡터(weight vector)를 찾아 간다. : winning node (or winner)
<b>Step 4</b>	각 레코드는 그리드 상의 winner와 그 이웃의 노드들을 동일한 방향으로 학습(업데이트)시킨다.
<b>Step 5</b>	이웃의 범위와 업데이트의 정도를 점차 줄임으로써 수렴 해를 얻는다.



# Kohonen SOM 군집분석 (3/3)

## ❖ 군집분석 결과의 예

- Jitter는 각 점에 있는 관측치를 일정 정도 흩으려 놓는 기능을 하여 각 좌표점에 얼마만큼의 관측치가 있는가를 보여줌.
- 코호넨 군집분석의 특징은 군집간의 관계를 2차원 그래프로 시각화(Visualization)하여 보여주는 것임.



## 분석결과

- $(0,0)$ ,  $(0,1)$ ,  $(0,2)$ ,  $(1,0)$ ,  $(1,1)$ ,  $(1,2)$ ,  $(2,0)$ ,  $(2,1)$ ,  $(2,2)$  등 9개 격자점을 중심으로 자료점들이 찍힘으로써 각 군집의 상대적 크기를 짐작할 수 있다.
- 플롯에서  $(2,0)$  격자 군집이 가장 크고 그 다음으로  $(0,0)$ ,  $(0,2)$ ,  $(2,2)$  격자 군집이 상대적으로 크게 나타나 있다.



# 군집분석 알고리즘 비교

## K-평균 군집화

- ✓ K-Means Clustering
- ✓ 타 알고리즘에 비해 연산시간이 적게 소요됨
- ✓ 군집 수 결정이 어려움
- ✓ 대규모 자료의 군집화 시 유용

## 코호넨 네트워크

- ✓ Kohonen Network
- ✓ 타 알고리즘에 비해 연산시간이 많이 소요됨
- ✓ 군집간의 상대적 거리계산이 용이함
- ✓ 군집 수 결정이 어려움

## 2단계 군집화

- ✓ TwoStep Clustering
- ✓ 대규모 자료의 군집화에 매우 효율적
- ✓ 1단계는 순차적으로 진행되기 때문에 개체들의 진입순서에 따라 결과가 상이하게 나타날 수 있으므로, 개체들의 순서가 무작위 (random) 이어야 함

# Case Study – 백화점고객 세분화 (1/4)

## 상황

국내 백화점 B사는 고객을 효율적으로 관리하기 위해서 특성이 비슷한 고객을 분류하고자 한다.

## 데이터

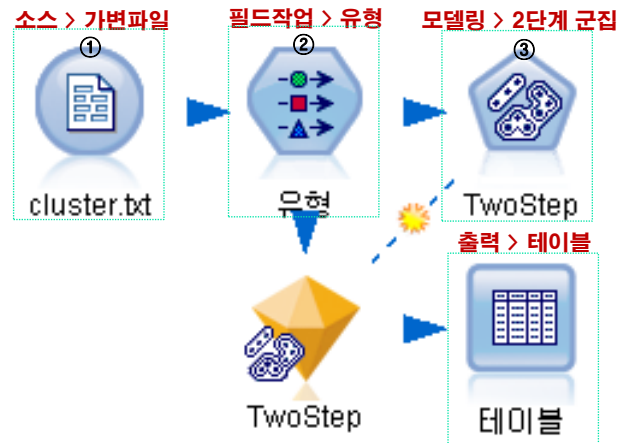
백화점 B사 고객 1,000명의 총 매출액, 누적방문빈도, 교차구매지수

## 분석 과정

① 데이터 준비 → ② 변수 지정 → ③ 2단계 군집분석

Data: cluster.txt

## 2단계 군집분석 과정



# Case Study – 백화점 고객 세분화 (2/4)

## ▷ 변수 목록

No.	변수 이름		변수 설명	변수 유형
	SPSS용	SAS용		
1	ID	ID	고객 고유번호	수치형
2	총 매출액	money	2년 동안 구매금액의 총합	수치형
3	누적방문빈도	visit	2년 동안 방문빈도의 총합	수치형
4	교차구매지수	cross	구매한 상품 종류의 수	범주형
5	평균구매주기	API	평균적으로 상품을 구매하는 주기(일)	수치형



# Case Study – 백화점 고객 세분화 (3/4)

## ▷ 2단계 군집분석 결과 (1/2)

- 2단계 군집모형의 결과 중 모형요약과 군집크기는 다음과 같이 확인할 수 있음.
- 2단계 군집모형 결과 자동으로 결정된 군집의 수는 3개임.



### ▪ 2단계 군집모형 > 모델 > 모형요약

모형 요약

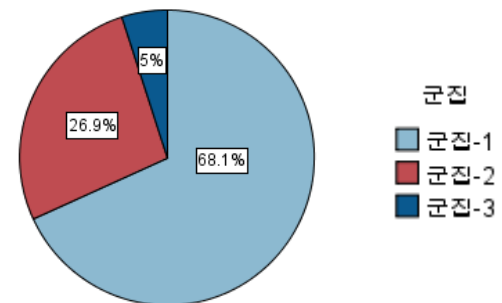
알고리즘	TwoStep
인력	3
군집	3

군집 품질



### ▪ 2단계 군집모형 > 모델 > 군집크기

군집 크기



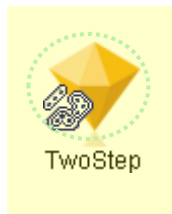
가장 작은 군집 크기	50 (5%)
가장 큰 군집 크기	681 (68.1%)
크기 비율: 가장 큰 군집에서 가장 작은 군집	13.62

# Case Study – 백화점 고객 세분화 (4/4)

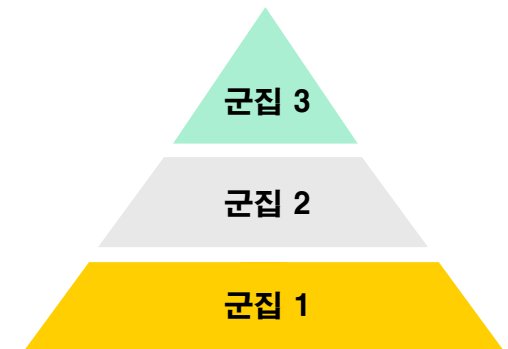
## ▷ 2단계 군집분석 결과 (2/2)

• K-평균 군집모형의 결과 중 입력변수의 군집 별 평균을 다음과 같이 확인할 수 있음.

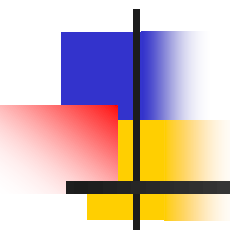
### ▪ 2단계 군집 모형 > 모델 > 군집



군집	군집-1	군집-2	군집-3
설명			
크기	68.1% (681)	26.9% (269)	5.0% (50)
입력	교차구매지수 3.00	교차구매지수 9.98	교차구매지수 17.30
	누적방문빈도 6.66	누적방문빈도 26.71	누적방문빈도 77.86
	총매출액 220,622	총매출액 1,056,663	총매출액 4,506,735



군집1 -> 군집2 -> 군집3 으로 갈수록 **교차구매지수, 누적 방문빈도, 총 매출액이 증가함**을 확인할 수 있음.



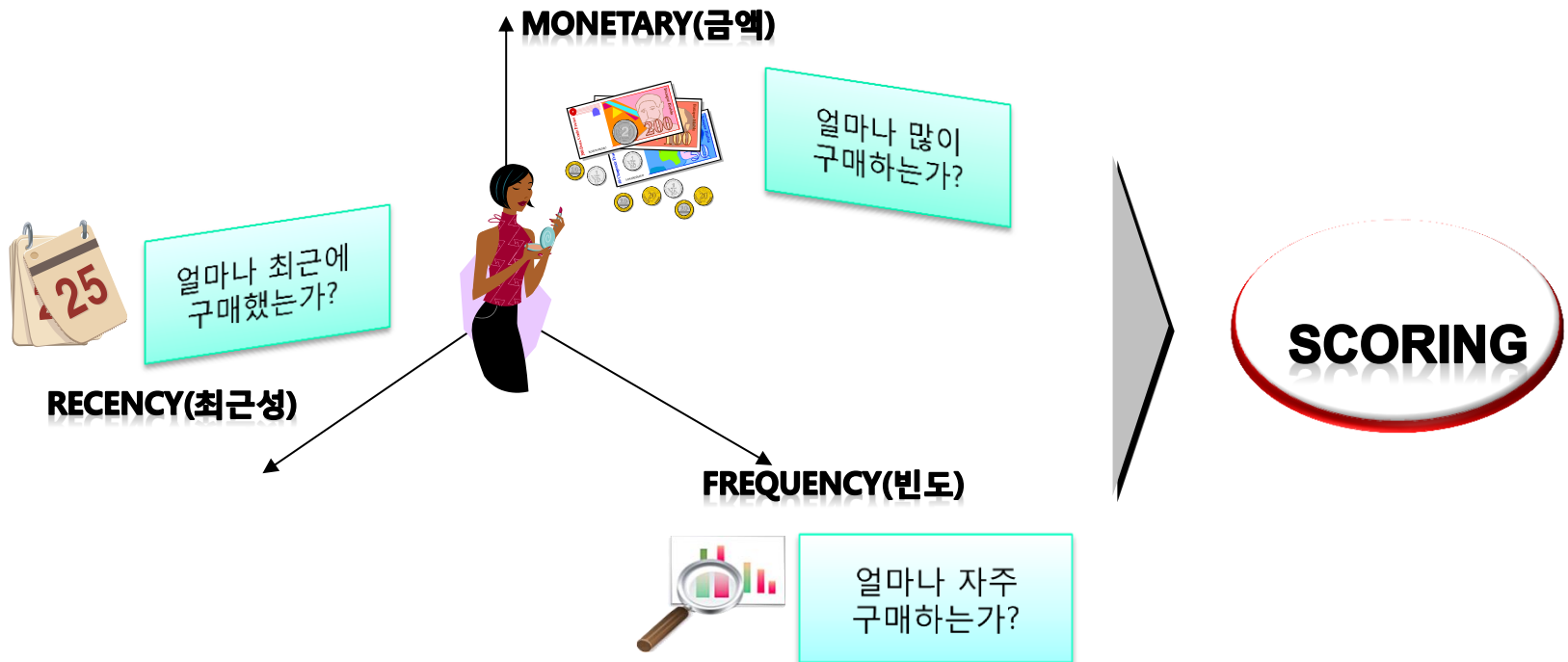
# RFM 분석

---

# RFM 분석이란?

## RFM analysis

- ✓ 고객을 세분화하는 방법으로 고객의 가치를 Recency(최근성), Frequency(빈도), Monetary(금액)의 기준에 따라 점수화하여 우량고객을 선별하는데 사용된다.
- ✓ 고객들의 가치를 판단하고, 이들에 대한 마케팅 효율을 높이며, 수익을 극대화 하도록 해주는 중요한 요소이다.



# RFM 분석 기초 (1/2)

## 거래내역데이터 통합하기

### <Transaction Data>

CardID	Date	Amount
1	20010820	229
1	20010628	139
1	20011229	229
2	20010727	49
2	20010202	169.99
2	20010712	299
2	20010202	34.95
2	20010907	99
2	20010513	49
3	20010922	99.99
3	20010502	5.99
3	20011101	49
3	20011016	69
4	20010812	84
4	20010328	69
4	20010403	24.99
4	20011210	19.99

### 거래내역데이터를 살펴보면,

- CardID가 1번인 고객은 총 3번의 거래를 하였으며 가장 최근의 거래일은 2001년12월29일이고, 3번의 구매에서 지출한 총 금액은 597이다.
- CardID가 2번인 고객은 총 6번의 거래를 하였으며 가장 최근의 거래일은 2001년09월07일이고, 6번의 구매에서 지출한 총 금액은 700.94이다.
- CardID가 3번인 고객은 총 4번의 거래를 하였으며 가장 최근의 거래일은 2001년11월01일이고, 4번의 구매에서 지출한 총 금액은 223.98이다.
- CardID가 4번인 고객은 총 4번의 거래를 하였으며 가장 최근의 거래일은 2001년12월10일이고, 4번의 구매에서 지출한 총 금액은 197.98이다.

# RFM 분석 기초 (2/2)

## 거래내역데이터 통합하기

<Transaction Data>

CardID	Date	Amount
1	20010820	229
1	20010628	139
1	20011229	229
2	20010727	49
2	20010202	169.99
2	20010712	299
2	20010202	34.95
2	20010907	99
2	20010513	49
3	20010922	99.99
3	20010502	5.99
3	20011101	49
3	20011016	69
4	20010812	84
4	20010328	69
4	20010403	24.99
4	20011210	19.99
.	.	.

<통합된 데이터>

CardID	Recency	Frequency	Monetary
1	2606	3	597
2	2719	6	700.94
3	2664	4	223.98
4	2625	4	197.98
.	.	.	.

통합

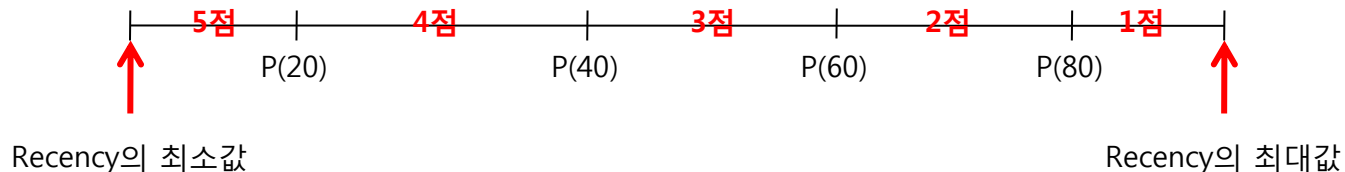
**1번 고객의 경우**, 현재날짜(2009년02월16일)와 최근거래날짜(2001년12월29일) 사이의 경과일수가 2606일이며, 3번의 거래에 총 597의 금액을 지출하였다. 이것을 하나의 레코드로 통합하였다.

# RFM 분석 알고리즘

## RFM 점수 계산법

1. 통합된 데이터의 Recency, Frequency, Monetary를 각각 순서대로 나열한 후, percentile을 가지고 동일크기의 레코드로 구간화 하여 각 구간에 같은 점수를 부여한다.

예를 들어, Recency의 최대 점수를 5점으로 설정하였다면 전체 Recency 값을 순서대로 나열한 후, 20percentile보다 작은 Recency 값을 갖는 레코드에는 5점을, 20percentile보다 크고 40percentile보다 작은 Recency 값을 갖는 레코드에는 4점을 주는 방법으로 1점까지 부여한다.



마찬가지로 Frequency와 Monetary도 위와 같은 방법으로 점수를 부여한다. Recency, Frequency, Monetary는 독립으로 가정하여 구한다. Frequency와 Monetary는 큰 값에 높은 점수를 부여해야 한다.

2. 앞서 점수를 부여하여 Recency Score, Frequency Score, Monetary Score를 생성하였다면, 각 점수에 중요도에 따라 가중값을 임의로 주어 RFM Score를 생성한다.

예를 들어, 3가지 항목에 같은 가중값 10을 주고 싶다면 다음과 같이 RFM 점수를 계산할 수 있다.

$$10 * \text{Recency Score} + 10 * \text{Frequency Score} + 10 * \text{Monetary Score} = \text{RFM Score}$$

# RFM 점수 계산

## RFM 점수 계산하기

percentile을 사용하여 동일크기 레코드가 할당되는 5개의 점수구간을 구하였다.



통합된 데이터를 가지고 다음과 같이 RFM 점수를 얻을 수 있다.

<통합된 데이터>

CardID	Recency	Frequency	Monetary
1	2606	3	597
2	2719	6	700.94
3	2664	4	223.98
4	2625	4	197.98
.	.	.	.

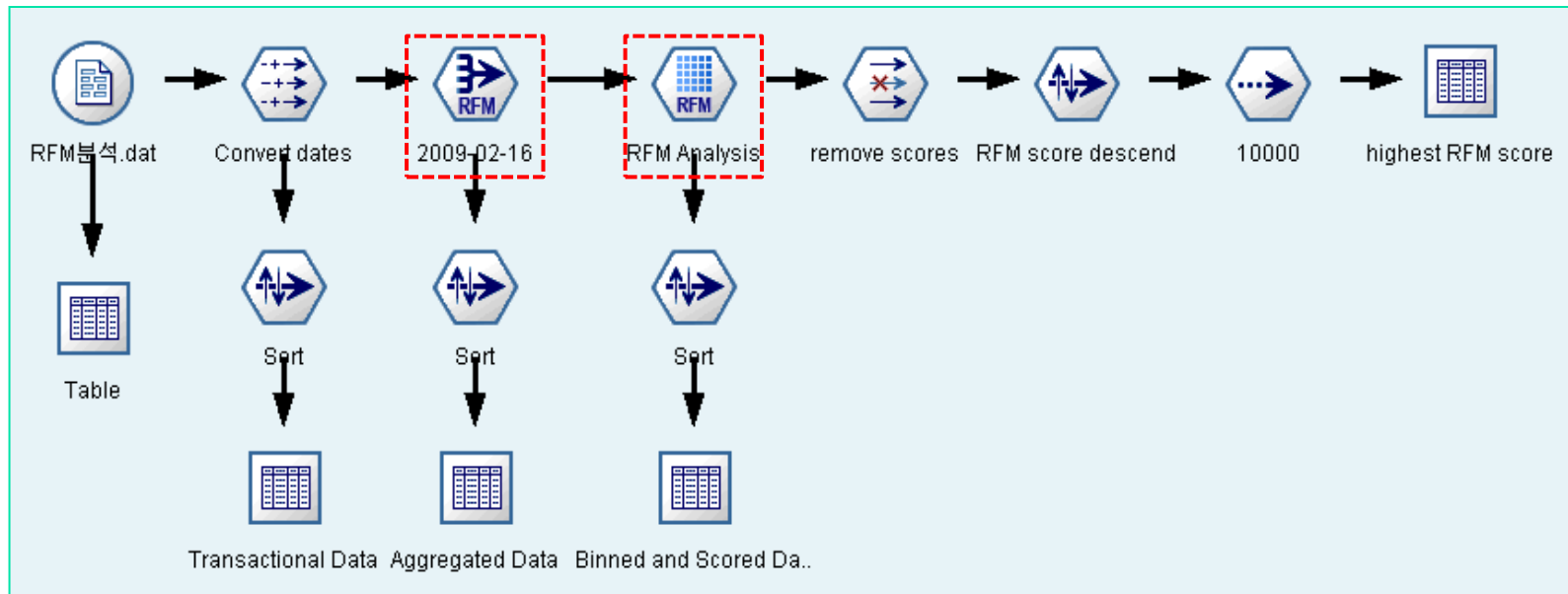


CardID	Recency Score	Frequency Score	Monetary Score	RFM Score
1	5	2	4	110
2	2	5	5	120
3	3	3	2	80
4	4	3	2	90
.	.	.	.	.



# RFM 분석 실습 스트림

거래내역데이터로 고객의 가치를 평점화하여 우량고객을 선별한다.



## ▪ RFM Aggregate 노드

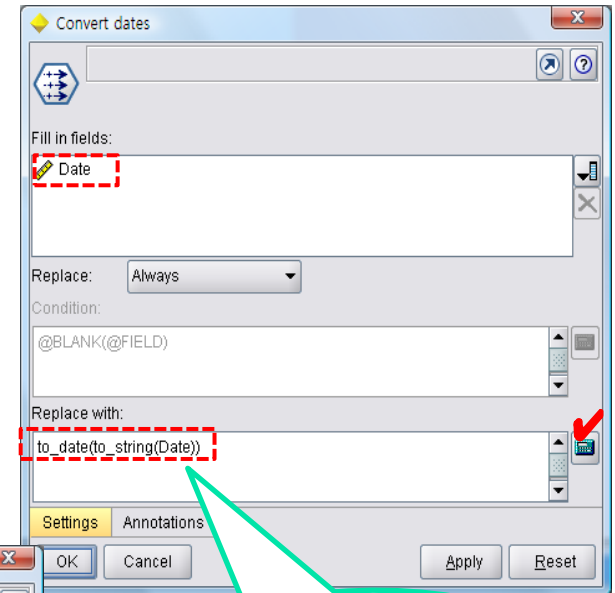
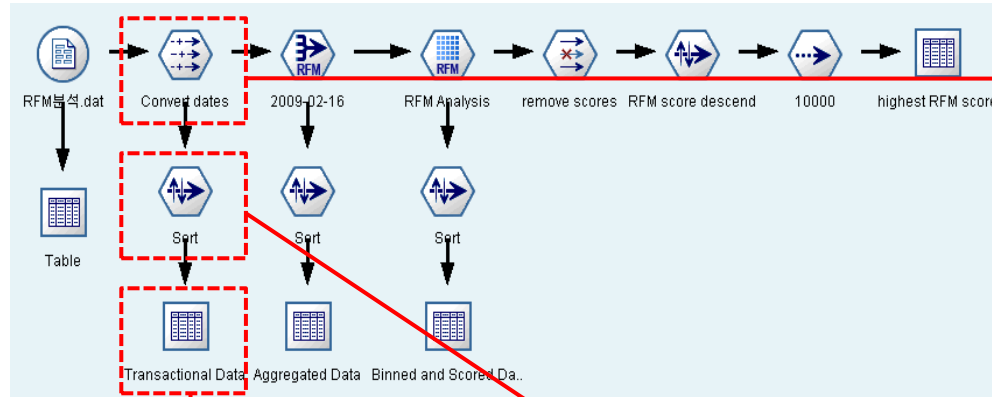
: 고객의 거래내역데이터(transaction data)를 고객ID를 기준으로 통합하는 기능을 한다.



## ▪ RFM Analysis 노드

: 고객의 구매정보가 요약된 Recency, Frequency, Monetary필드를 사용하여 점수화한 후 RFM score를 산출한다.

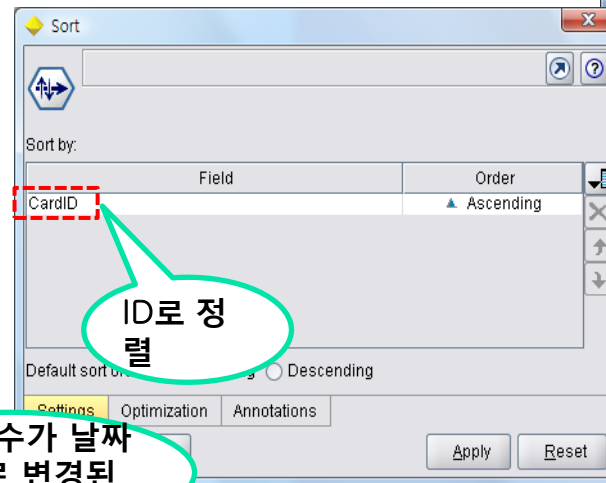
# Step1. 데이터 설명 및 핸들링 (1/2)



처음에 데이터를 불러 오면 Date변수를 숫자 형으로 인식하므로 문자형으로 바꾼뒤 날짜로 변환.

Transactional Data (3 fields, 69,215...)

	CardID	Date	Amount
1	C0100000199	2001-08-20	229.000
2	C0100000199	2001-06-28	139.000
3	C0100000199	2001-12-29	229.000
4	C0100000343	2001-02-02	34.950
5	C0100000343	2001-07-12	299.000
6	C0100000343	2001-09-07	99.000
7	C0100000343	2001-05-13	49.000
8	C0100000343	2001-02-02	169.990
9	C0100000343	2001-07-27	49.000
10	C0100000375	2001-10-16	69.000
11	C0100000375	2001-09-22	99.990
12	C0100000375	2001-05-02	5.990
13	C0100000375	2001-11-01	49.000
14	C0100000482	2001-08-12	84.000
15	C0100000482	2001-03-28	

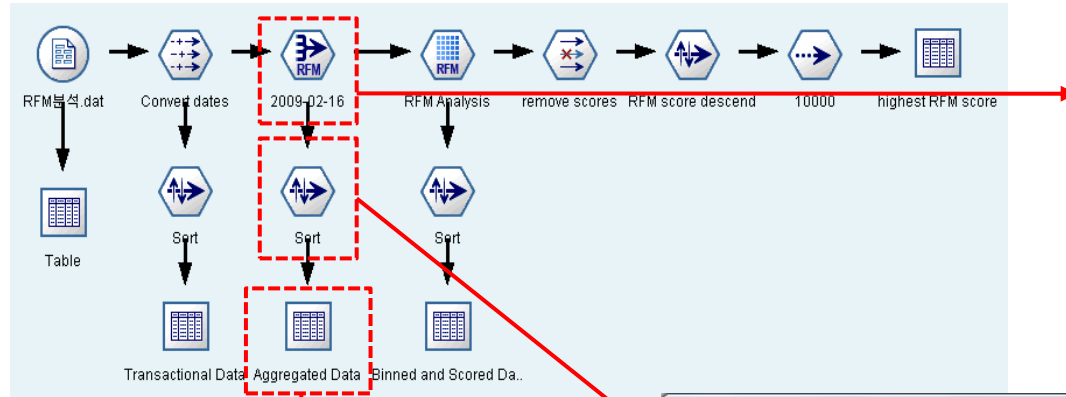


ID로 정렬

Date 변수가 날짜 형식으로 변경된 것을 확인

CLEM 언어	설명
to_date	항목을 날짜로 변환
to_string	항목을 문자열로 변환

# Step1. 데이터 설명 및 핸들링 (2/2)



Aggregated Data (4 fields, 12,589 records)

	CardID	Recency	Frequency	Monetary
1	C0100000199	2606	3	597.000
2	C0100000343	2719	6	700.940
3	C0100000375	2664	4	223.980
4	C0100000482	2625	4	197.980
5	C0100000689	2609	2	428.000
6	C0100000789	2606	3	777.000
7	C0100000915	2615	1	49.000
8	C0100001116	2737	6	942.970
9	C0100001139	2716	4	339.490
10	C0100001156	2670	2	528.000
11	C0100001244	2771	3	339.930
12	C0100001405	2801	2	153.990
13	C0100001916	2638	5	371.980

Table Annotations

Sort

Sort by:

Field	Order
CardID	Ascending

Sort order: ☒ Ascending ☐ Descending

Apply Reset

2009-02-16

Calculate Recency relative to: ☒ Fixed date 2009-02-16 ☐ Today's date

☐ IDs are contiguous

ID: ☒ CardID ☒ Date ☒ Amount

Value: ☒ Amount

New field name extension:  Add as: ☐ Suffix ☒ Prefix

☐ Discard records with value below:  1.0

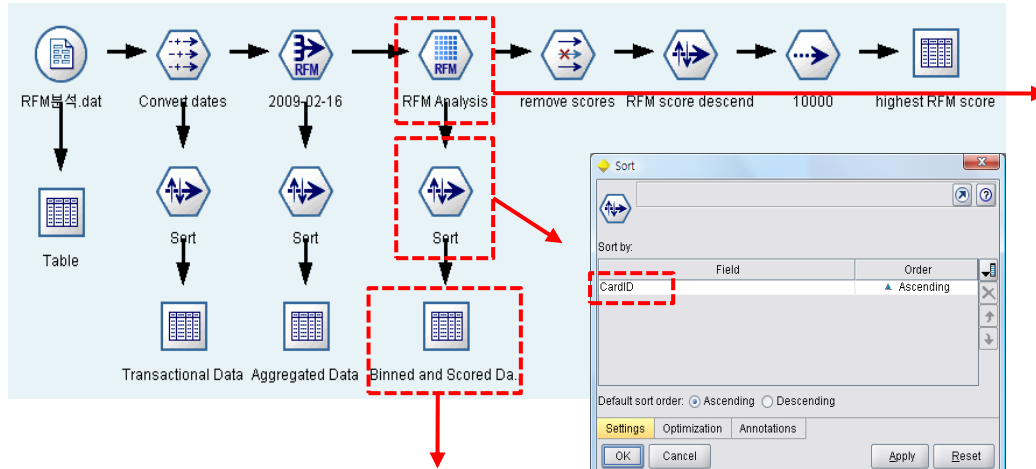
☐ Only include recent transactions:

Apply Reset

데이터를 통합하기 위해 ID, Date, Amount에 해당하는 변수를 지정함.

거래데이터가 ID를 기준으로 통합되어 최근거래부터 현재까지 거래일수(Recency), 거래빈도(Frequency), 총거래금액(Monetary) 변수가 생성됨.

# Step2. RFM분석 (1/2)



점수화 할 변수를 선택함

RFM Analysis

Thresholds may be edited and will only be recomputed if necessary

Recency: ☒ Recency ✓

Frequency: ☒ Frequency ✓

Monetary: ☒ Monetary ✓

	Number of bins	Weight
Recency:	5	10.0
Frequency:	5	10.0
Monetary:	5	10.0

Ties: ☐ Add to next ☒ Keep in current

Bin thresholds: ☐ Always recompute ☒ Read from Bin Values tab if available

☒ Add outliers to end bins

Settings Bin Values Annotations

OK Cancel Apply Reset

최대 점수와 가중값을 지정

Binned and Scored Data (8 fields, 12,589 records)

	CardID	Recency	Frequency	Monetary	Recency Score	Frequency Score	Monetary Score	RFM Score
1	C0100000199	2606	3	597.00	5	2	4	110.000
2	C0100000343	2719	6	700.94	2	5	5	120.000
3	C0100000375	2664	4	223.98	3	3	2	80.000
4	C0100000482	2625	4	197.98	3	3	2	80.000
5	C0100000689	2609	2	428.00	5	1	1	70.000
6	C0100000789	2606	3	777.00	5	2	4	110.000
7	C0100000915	2615	1	49.00	5	1	1	70.000
8	C0100001116	2737	6	942.97	2	5	5	120.000
9	C0100001139	2716	4	339.49	2	3	3	80.000
10	C0100001156	2670	2	528.00	3	1	4	80.000
11	C0100001244	2771	3	339.93	1	2	3	60.000
12	C0100001405	2801	2	153.99	1	1	1	30.000

Table Annotations

OK

$$2(R) \cdot 10 + 5(F) \cdot 10 + 5(M) \cdot 10 = 120$$

각 부분별 점수가 계산되고 설정한 가중값으로 계산된 RFM점수가 생성됨.

RFM Analysis

Thresholds may be edited and will only be recomputed if necessary

Binned field: Recency

Binning settings have changed. Select the Read Values button to update the grid.

Bin	Lower	Upper
1	> 2755	<= 2968
2	> 2684	<= 2755
3	> 2643	<= 2684
4	> 2618	<= 2643
5	>= 2605	<= 2618

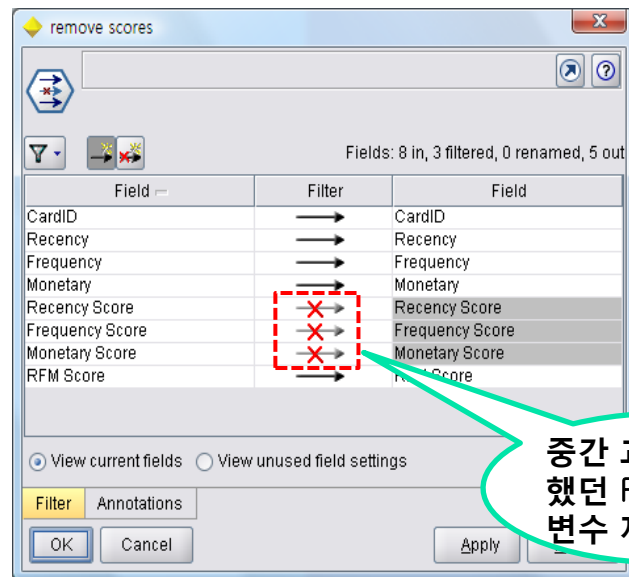
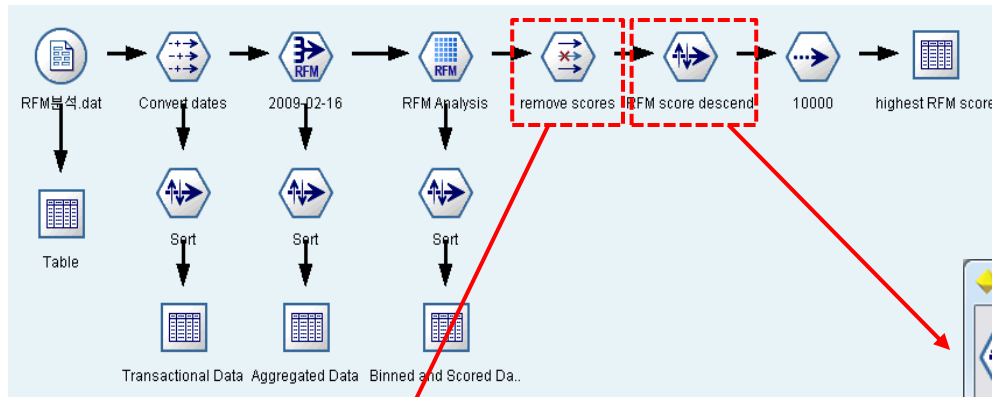
Read Values

Settings Bin Values Annotations

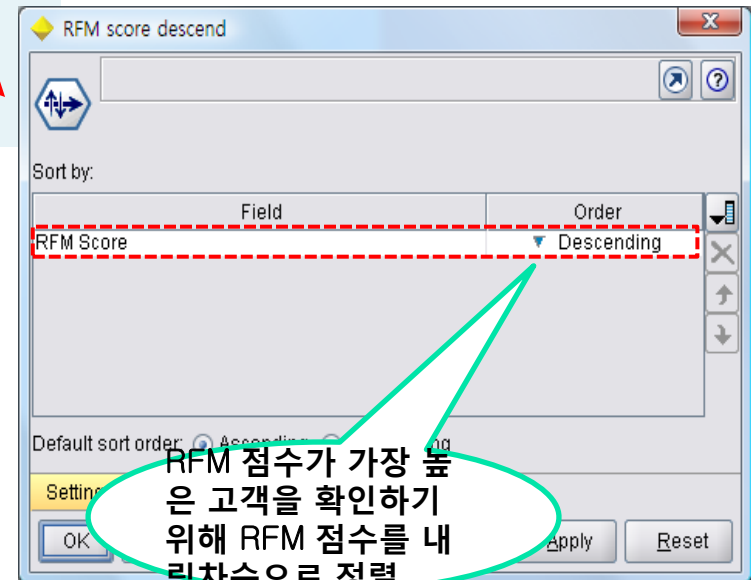
OK Cancel Apply Reset

Recency 점수의 기준을 확인

## Step2. RFM분석 (2/2)

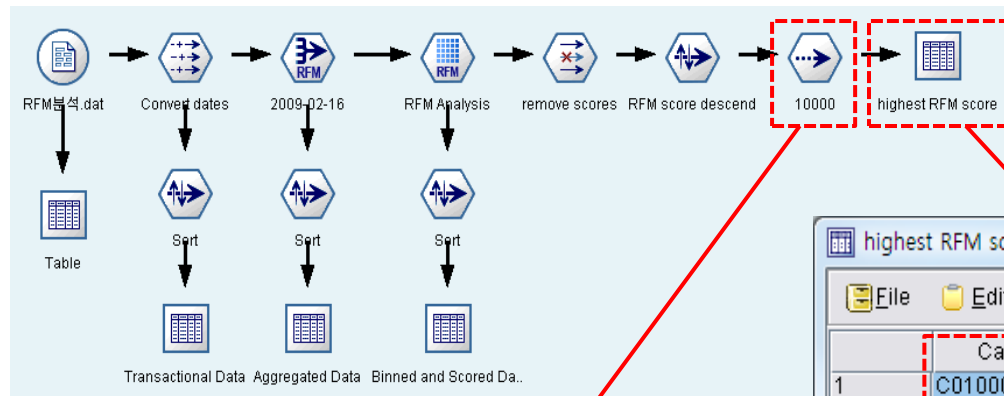


중간 과정에 필요했던 R/F/M 점수 변수 제거



RFM 점수가 가장 높은 고객을 확인하기 위해 RFM 점수를 내림차순으로 정렬

# Step3. RFM 분석 결과



RFM 점수가 큰 고객 순으로 정렬됨.

highest RFM score (5 fields, 10,000 records) #1

	CardID	Recency	Frequency	Monetary	RFM Score
1	C0100075034	2607	29	718.680	150.000
2	C0100559670	2614	29	600.690	150.000
3	C0102085654	2605	22	763.310	150.000
4	C0105096521	2611	23	790.760	150.000
5	C0100154614	2612	13	975.400	150.000
6	C0100154629	2605	10	671.890	150.000
7	C0102064750	2615	26	958.780	150.000
8	C0104204270	2609	12	1671.840	150.000
9	C0103403450	2615	28	979.710	150.000
10	C0102060029	2610	8	981.920	150.000
11	C0104975050	2613	25	880.240	150.000
12	C0101347480	2612	17	1701.360	150.000
13	C0105117594	2610	29	998.740	150.000
14	C0102934051	2618	29	1830.670	150.000
15	C0105119123	2609	23	1768.820	150.000
16	C0106139786	2610	29	905.180	150.000

Annotations

RFM 점수가 높은 10000명의 고객만 추출

RFM 점수를 통하여 고객 세분화를 할 수 있음.